DOCUMENT RESUME

ED 197 712 IR 009 098

AUTHOR	Roistacher, Richard C.: And Others
TITLE	A Style Manual for Machine-Readable Data Files and Their Documentation.
INSTITUTION	Bureau of Social Science Research, Inc., Washington, D.C.
SPONS AGENCY	National Criminal Justice Information and Statistics Service (Dept. of Justice/LEAA), Washington, D.C.
REPORT NO	NCJ-62766: SD-T-3
PUB DATE	Jun 80
GRANT	78-SS-AX-0028
NOTE	82p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402 (1980 311-379/1412).
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Databases: *Data Processing: Documentation: Information Retrieval: *Information Storage: *Methods: *Standards: *Statistical Data
IDENTIFIERS	*Machine Readable Data

ABSTRACT

This manual presents detailed descriptions and examples of standards and techniques for formatting and documenting machine-readable data files. The descriptions of syntax and stylistic elements are independent of whether documentation is produced manually or as a machine-readable text file. The manual also discusses rules of good practice for producing and documenting machine-readable data files. (Author/RAA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED197712

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A STYLE MANUAL FOR MACHINE-READABLE DATA FILES AND THEIR DOCUMENTATION

Report Number SD-T-3, NCJ-62766

June 1980

By Richard C. Roistacher
Bureau of Social Science Research
Washington, D.C.

With contributions from

Sue A. Dodd
Institute for Research in Social Science
University of North Carolina, Chapel Hill

Barbara B. Noble
Bureau of Social Science Research
Washington, DC

Alice Robbin
Data and Program Library Service
University of Wisconsin-Madison

This report was supported by Grant No. 78-SS-AX-0028 awarded to the Bureau of Social Science Research by the Statistics Division, National Criminal Justice Information and Statistics Service, Law Enforcement Assistance Administration (now the Bureau of Justice Statistics), U. S. Department of Justice, under the Omnibus Crime Control and Safe Streets Act of 1968, as amended. The project which produced this report was directed for the Bureau of Social Science Research by Richard C. Roistacher and monitored for BJS by Marianne W. Zawitz.

Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position on policies of the U. S. Department of Justice.

BJS authorizes any person to reproduce, publish, translate, or otherwise use all or part of the material in this publication.

IR 009098

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402

OCT 6 1980

U. S. Department of Justice
Bureau of Justice Statistics

Benjamin H. Renshaw, Acting Director

Charles R. Kinderman, Acting Director
Statistics Division

TABLE OF CONTENTS

Abstract i
List Of Examples v
Acknowledgments 1
Chapter 1: Introduction 2
Organization Of This Manual 3
Format Of This Manual 4
Documentation Examples 4
Page Formats 4
Headings 4
Chapter 2: Preliminaries 6
Bibliographic Identity 6
Title 7
Subtitle 8
Authorship 9
Author Responsibility Statement 9
Edition 9
Edition Responsibility Statement 9
Producer 10
Distributor 10
Title Page 10
Pagination And Tables Of Contents 19
Chapter 3:
History Of The Originating Project 22
Instrumentation 22
Bibliography 23
Chapter 4: File Processing Summary 27
Archive File Description 27
Confidentiality Procedures 27
Chapter 5:
Data Dictionary Description And Listing 31
Dictionary Information 31
Wording Of Questions 31
Variable Names Or Numbers 31
Reference Numbers 32
Variable Labels 32
Explanatory Text 32
Response Text 33
Universe Definition 33
Description Of Data Dictionary 33
Chapter 6: Appendices 40
Definitions 40
Errors And Problems 42
Extended Response Categories 45
Original Data Collection Instrument 47
Field Or Laboratory Procedures 47
Dictionary Format 47
Description Of Physical Shipment 50

Order Form 52
Chapter 7:
Documenting Machine-readable Data: A Checklist 54
Chapter 8:
Technical Standards For Machine-readable Data Files .66
Tape Recording Standards 66
Data Types 67
Missing Data 68
File Organization 69
Data Items 69
Record Type Identification 69
Record Identification Items 70
Dates 71
Standardization Of Data Codes 71
Documentation 72
Minimal Documentation 72
Tape Table Of Contents 72
Minimal Codebook 75
Frequency Tables 75

LIST OF EXAMPLES

1. User's Guide Title Page 13
2. Abstract 17
3. User's Guide Table Of Contents 20
4. Data Dictionary Table Of Contents 21
5. Project History 24
6. File Processing Summary 29
7. Description Of Data Dictionary Listing 34
8. Data Dictionary Listing 36
9. Appendix: Definitions Of Terms 41
10. Appendix: Errors And Problems In The Data 43
11. Appendix: Extended Code Categories 46
12. Appendix: OSIRIS III Type 4 Dictionary 48
13. Appendix: Data Shipment Description 51
14. Data Order Form 53
15. Tape Volume Table Of Contents 74

ABSTRACT

This manual presents a detailed description of standards and techniques for formatting and documenting machine-readable data files. The manual includes extensive examples of documentation. The descriptions of syntax and stylistic elements are independent of whether documentation is produced manually or as a machine-readable text file. The manual also discusses rules of good practice for producing and documenting machine-readable data files.

This manual was developed under the following grants from the National Criminal Justice Information and Statistics Service (now the Bureau of Justice Statistics), Law Enforcement Assistance Administration (LEAA), U. S. Department of Justice: 77-SS-99-6003, 77-SS-99-6021, 78-SS-AX-0028.

ACKNOWLEDGMENTS

Many people and institutions contributed to this manual. Funding was provided by the National Criminal Justice Information and Statistics Service (now the Bureau of Justice Statistics), Law Enforcement Assistance Administration, U. S. Department of Justice, under grants 77-SS-99-6003, 77-SS-99-6021, and 77-SS-AX-0028. Materials and methods have been appropriated from the Inter-university Consortium for Political and Social Research, DUALabs, the U. S. Bureau of the Census, and the LEAA Research Support Center, among others.

Many contributions have been made by the advisory committee, consisting of:

Sue A. Dodd, Institute for Research in Social Science, University of North Carolina, Chapel Hill;

Carolyn Geda, Inter-university Consortium for Political and Social Research, University of Michigan, Ann Arbor;

Charles Gellert, Machine Readable Data Division, National Archives and Records Service, Washington, DC;

Warren Glimpse, Data User Services Division, Bureau of the Census;

Barbara B. Noble, Bureau of Social Science Research, Washington, DC;

A. Elizabeth Powell, National Academy of Public Administration;

Debra Powell, DUALabs, Rosslyn, VA;

Alice Robbin, Data and Program Library Service, University of Wisconsin-Madison;

Judith Rowe, Princeton University Computing Center, Princeton, NJ.

CHAPTER 1: INTRODUCTION

The purpose of this manual is to describe the production of a fully documented machine-readable data file (MRDF). A fully documented machine-readable data file as discussed here consists of two kinds of objects carrying two kinds of information: a tape or other medium carrying the machine-readable data themselves and a book of documentation, which tells how the data are to be read and interpreted.

MRDF documentation consists of information which describes the file's identity, organization, contents, physical characteristics, and relation to computer hardware and software. Items of documentation have been called codebooks, tape layouts, data dictionaries, program manuals, etc. This manual describes the format of a comprehensive manual of documentation, which will be called a "user's guide." This term has been chosen for its generality, simplicity, aptness, and conformity with Federal Information Processing Standards (FIPS) for documentation.

The definition of "machine-readable" data changes daily as technology makes it possible for an ever-increasing variety of information to be captured in machine-readable form. This manual concerns the documentation and formatting of machine-readable data in a limited sense, i.e., as files of numeric information in rectangular or tree-structured form. This manual does not cover the documentation of files of free-format text.

Although most of the examples shown here are drawn from the social sciences, the manual is not restricted to data from any particular discipline. All machine-readable data in the form of numeric or coded records to be preserved for later analysis should be documented in the same fashion. It matters little whether a file consists of responses to a questionnaire, income figures for counties, or temperatures and barometric pressures. A file's format and documentation style is affected by whether the file represents a data matrix, a tree, a hierarchy, a graph, or a multidimensional table. However, its format and documentation are not affected by the nature of the data which are stored in the matrix, tree, hierarchy, or table.

This manual is designed to be used by data producers and archivists. The manual presents a style of documentation appropriate for all files designed to be used as continuing resources. Whenever possible, data producers should document their MRDFs in the full style described here. This manual also describes a minimal set of documentation for a MRDF. No file, however temporary, can be considered adequately documented with less than the minimal documentation described here.

The preparation of documentation is a task for the data producer. While an archive may polish the original

documentation into a user's guide, the archive should not have to do any substantive documentation itself. In particular, there should be no need for the archive to apply to the producer for materials not provided with a file's documentation.

Until recently, little effort was made to catalog machine-readable data files. However, there is now an increasing emphasis on preserving data for secondary analysis. If machine-readable data files are to be long-term research resources, then they must adhere to bibliographic standards of quality and identity. The requisite standard of quality will be achieved when data producers properly identify their data files with sufficient bibliographic elements, which in turn can be processed by librarians and converted into catalog records and integrated into existing information systems.

The requisite standards of identity will be achieved only if data users adhere to common standards of citation. Once a machine-readable data file has been given a bibliographic identity, then it must be used and cited with that identity. Only in this way will proper credit be given to data producers and primary analysts. In particular, users who pass on altered copies of data are responsible for seeing that the bibliographic identity is maintained, but that a new edition is recorded when appropriate.

The addition of machine-readable data files to union lists and library collections will increase their circulation and scientific utility. However, it will become increasingly necessary for data users to refrain from making unnecessary alterations either in the data file or in its bibliographic descriptor.

Organization Of This Manual

This manual consists of eight chapters. Five of these chapters cover the five major sections of a user's guide to a machine-readable data file. These are:

1. Preliminaries: Bibliographic attributes, title and title page construction, pagination, and headings;

2. A history of the project which produced the MRDF, describing the evolution of the data from the point of collection until conversion to machine-readable form;

3. A summary of the MRDF's data-processing history;

4. A dictionary listing of the data items in the file (the "codebook" proper);

5. A set of appendices containing glossaries, error listings, bibliographies, and other information.

The other chapters consist of this introduction, a checklist for documentation, and a set of technical standards for MRDF.

Format Of This Manual

Much of this manual consists of heavily annotated sections of a MRDF user's guide. Thus the manual itself is in an editorial style appropriate for a MRDF user's guide. While this manual (and much MRDF documentation) was produced with a computer text formatting system, the documentation style described here does not require the use of such a system.

Documentation examples. This manual uses examples of documentation from several academic and government sources. Many of these excerpts have been modified to conform to the style expounded here. None of the examples in this manual should be construed as actual documentation or as an accurate bibliographic citation for an existing machine-readable data file.

Page formats. If documentation is to be machine-readable, it should be designed for printing on a line printer in a format suitable for reproduction on standard size paper. Text should be designed for printing on one or both sides of an 8-1/2 x 11-inch sheet of paper, with a 1-1/2 inch left margin and 1-inch top, right, and bottom margins. Titles extend 1/2 inch into the top margin; footers extend 1/2 inch into the bottom margin. This manual is itself an example of machine-readable documentation and is formatted to the standards given here. Text is single spaced at six lines per vertical inch and 10 characters per horizontal inch.

Headings. This manual uses the four levels of heading described in the Publication Manual of the American Psychological Association, Second Edition. The APA style was chosen because it seemed generally appropriate and because it is well documented. Since most scholarly editors prefer a consistent publication style, regardless of origin, to an inconsistent approximation of their own, the APA style can be recommended for documenting machine-readable data for any constituency. The four heading types are:

AN UPPER-CASE CENTERED MAIN HEADING

A Centered and Underlined Main Heading

A Flush Underlined Capitalized Side Heading

An indented paragraph heading. The higher-level headings are either in upper case or capitalized. The indented paragraph heading is in sentence format and is immediately followed by text.

Capitalization and underlining or italicization are as shown. The fourth level heading runs directly into the text. When only three levels of headings are to be used throughout, it is acceptable to use headings 1, 3, and 4 if desired, rather than 1, 2, and 3.

CHAPTER 2: PRELIMINARIES

A machine-readable data file's bibliographic identity is provided by six kinds of information:

1. Information which identifies the MRDF and prepares it for integration into existing manual or automated information systems. This information consists of bibliographic elements which can be processed by librarians, converted to catalog records, and integrated into bibliographic storage and retrieval systems.

2. Information which describes the contents of a MRDF. This information is contained in a data abstract, which also can be automated and integrated into existing information systems.

3. Information which classifies a MRDF is contained in a set of descriptors or keywords which will facilitate the retrieval of a group of MRDFs on the same or similar subjects.

4. Information required to access a MRDF includes a description of the physical characteristics of the file and its relation to the computer hardware and software.

5. Information necessary to analyze the MRDF describes the data items in the file, the methods used to create the file, and the file's linkage to methodologically similar data files. This information is usually presented in a "data dictionary listing" or "codebook."

6. Information necessary to accession or archive the MRDF, to evaluate its quality, and to prepare it for future use.

This chapter provides guidance on how to identify, describe, and classify a machine-readable data file. Rules are suggested and illustrated with examples. The rationale for these rules is to help primary analysts (who are often the data producers) to receive proper credit and recognition for their contribution to research and to policy and program evaluation. Proper title pages, bibliographic citations, and abstracts will increase the opportunities for such recognition and will also provide essential information to the users of information systems.

Bibliographic Identity

The bibliographic identity of a MRDF is composed of the following elements:

1. Title
2. Subtitle
3. Authorship
4. Author responsibility statement
5. Edition
6. Edition responsibility statement
7. Producer
8. Distributor

These elements are discussed in detail below.

Title. A title for a MRDF should be descriptive of the contents of the data it is describing. If there is one main theme or focus, then this should be mentioned in the title. For example:

JUVENILE DETENTION AND CORRECTIONAL FACILITY CENSUS OF 1971

If the MRDF contains information on many different topics, none of which appear to dominate, then a broader or more general subject approach may be taken. For example:

NATIONAL OPINION RESEARCH CENTER 1974 AMALGAM SURVEY

If the MRDF are part of a predictable series--occurring at definite time intervals, such as the U. S. Census of Population, the Current Population Surveys, or the American National Election Surveys, then the titles should be structured in a consistent pattern throughout the life of the series. For example:

CURRENT POPULATION SURVEY: ANNUAL DEMOGRAPHIC FILE, 1968
CURRENT POPULATION SURVEY: ANNUAL DEMOGRAPHIC FILE, 1969
CURRENT POPULATION SURVEY: MULTIPLE JOB HOLDING, MAY 1972
CURRENT POPULATION SURVEY: ADULT EDUCATION PARTICIPANTS, MAY 1972

In cases where there is more than one data collection (or reporting period per year), the month or the seasons should become part of the title. For example:

SURVEY OF CONSUMER ATTITUDES AND BEHAVIOR, SPRING 1976
SURVEY OF CONSUMER ATTITUDES AND BEHAVIOR, SUMMER 1976

If the geographic focus is unique, it should be included in the title. For Example:

ARGENTINA DOMESTIC VIOLENCE AND ECONOMIC DATA, 1955-1972

For MRDF that are part of an ongoing collection (e.g., public opinion polls or surveys that are collected at unpredictable intervals and with varying subject focus), the following title structure may be used:

a. Organizational name of producer;

b. Chronological date of data collection;

c. Geographical focus (if unique);

d. Descriptive content phrase;

e. Series or study number.

This arrangement allows large collections of data from the same source to be grouped alphabetically for easy reference. For example:

HARRIS 1965 COLLEGE STUDENT SURVEY, No. 1431
HARRIS 1965 CRIME AND CIVIL RIGHTS SURVEY, No. 1561
HARRIS 1965 NEW JERSEY GUBERNATORIAL ELECTION SURVEY, No.. 1560
HARRIS 1965 VIETNAM TELEPHONE SURVEY, No. 1546

Subtitle. A subtitle is a secondary title used to amplify or state certain limitations of the main title. For example:

DETROIT AREA STUDY 1967: CITIZENS IN SEARCH OF JUSTICE
DETROIT AREA STUDY 1968: BLACK ATTITUDES IN DETROIT

Subtitles may also be used to indicate a special population sample or subfile. For example:

CARNEGIE COMMISSION NATIONAL SURVEYS OF HIGHER EDUCATION 1969: FACULTY STUDY SUBSAMPLE

COMPARATIVE STATE ELECTION PROJECT 1968: FEDERAL DISTRICT SUBFILE

Several things should be avoided when creating a title for MRDF, including:

1. Avoid beginning a title with articles.

2. Avoid beginning a title with numerics (e.g., 1972 County and City Data book).

3. Avoid using acronyms or data set names. If acronyms are used, their meaning should be spelled out and the acronym enclosed in brackets. For example:

WORLD EVENT/INTERACTION SURVEY (WEIS)

Authorship. Give the full name of author(s) and/or corporate body responsible for the design of the survey instrument or other source of intellectual content of the MRDF.

Author responsibility statement. This statement is used to indicate the relationship of the work to the person(s) named in the author heading that otherwise would not be known. For example:

CONDUCTED FOR THE NATIONAL DATA PROGRAM FOR THE SOCIAL SCIENCES AT THE NATIONAL OPINION RESEARCH CENTER, 1978.

Edition. An edition statement is not necessary if the MRDF is being produced for the first time. However, if there have been substantive changes in the MRDF since its creation, then this statement should be used. Separate revisions of the file's documentation or subsequent printings of the documentation independent of changes in in the MRDF do not constitute a new edition of the data file.

An edition for MRDF occurs when any one of the following conditions are fulfilled by a primary analyst, data producer, or designated archive:

1. Any additions or deletions of the original machine-readable text;
2. Any additions or deletions of data elements or variables;
3. Any recoding or reformatting of the file;
4. Any change in the number of logical records;
5. Any change in the programming language.

The phrasing of the edition statement is customarily abbreviated. For example:

FORTRAN IV ed.	ICPSR 2nd ed.
REV. 1972 ed.	SPSS TEST ed.

Edition responsibility statement. This statement is used to indicate the purpose or organization responsible for the MRDF edition, including the purpose or intended application of the new edition. For example:

NCES Rev. ed. Revised by Nancy Moxley for education

and classroom use.

Producer. The producer of a MRDF is the person or organization with the financial or administrative responsibility for producing the computerized file. Such responsibility includes that for the various tasks associated with collecting data or information and converting the information into a computerized format. These tasks may be contracted out or be performed by more than one party. Financial responsibility alone does not define the producer of a MRDF; if the tasks associated with MRDF production are contracted or carried out by more than one party, the arrangements should be cited in the file's documentation.

If no other individual or corporate author is identified, then it is assumed that the producer is also responsible for the file's intellectual content. The producer statement includes the name of the organization (or person), the location of the organization, and the date of production.

The production date for MRDF is interpreted as the date the file became operational in a computerized form and available for analysis, processing, and possible release to the general public. Unlike books and the act of publishing, the production of a MRDF does not always coincide with its release to the general public. In many cases, a MRDF may be produced in one year, analyzed and reported in another, and released to the public several years later by a party other than a producer (see the definition of distributor). In cases where the production date has to be approximated, then the date of data collection may be used as a point of reference.

Distributor. A MRDF distributor is defined as the organization designated by the author or producer to generate copies of a particular file, including any necessary editions or revisions. If a distributor is not cited, then it is assumed that the author or producer is fulfilling this function. A distributor statement includes the organization or (person) name, location (or address) of the distributor, and date of distribution.

The distribution date for a MRDF is the year the file became available for distribution to the general public, usually through an established agency.

Once defined, these bibliographic data elements can be used to create a title page and a bibliographic citation. They may also be used in the creation of a data abstract.

Title Page

The title page is the major source of bibliographic information for MRDF. Care should be taken to provide enough information to identify uniquely the file being described. The

title page should also include sufficient information on any major contributors to the file and any relevant relationships associated with its production and distribution.

Suggested bibliographic elements for a title page include: title, subtitle (if appropriate), personal or corporate authorship, and appropriate responsibility statements associated with the data production, including source of funding, organizational name of data producer, academic affiliation or address of producer, the production date, an edition statement (if appropriate), and organization name and address of distributor (if different from the producer).

If the production or collection date of the data varies from the publication date of the file's documentation, this latter date should be flagged accordingly:

> User's Guide Prepared by LEAA Data Archive and Research Support Center, 1978, with support from LEAA Grant 77-SS-99-6003.

Edition statements for the MRDF must also be distinguished from the edition statements associated with the file's documentation. The MRDF edition statement would follow the production statement on the title page; and the edition statement pertaining to the file's documentation would appear in close proximity to the date of the documentation's publication, normally placed at the bottom of the title page. Edition statements pertaining to the file's documentation should also be flagged. For example:

Edition statement for MRDF: LEAA rev. 1975 ed.

Edition statement for MRDF documentation: LEAA User's Guide rev. 1975 ed.

An outline of bibliographic elements for a title page, which may be used as a guide, follows. (Item numbers refer to the bracketed numbers in the next example.)

1. Title
2. Subtitle (if appropriate)
3. Special sponsorship or funding source
4. Producer's name
5. Date of production or collection
6. Authorship
7. Address
8. Edition statement (if appropriate) for data
9. Distributor's name, address, and telephone number
10. Acknowledgement of organization/funding source responsible for publishing the related documentation (use only if different from the producer or distributor of the data file)
11. Date of the documentation's publication
12. Edition statement (if appropriate) for documentation

Example 1 (page 13) shows a title page for a user's guide.

EXAMPLE 1: USER'S GUIDE TITLE PAGE

JUVENILE DETENTION AND CORRECTIONAL FACILITY CENSUS OF 1971 <1>

User's Guide for the Machine-Readable Data File <2>

Produced by

U. S. Bureau of the Census <4>
Washington, D.C.
1971 <5>

for

National Criminal Justice Information and Statistics Service <6>
Law Enforcement Assistance Administration
U. S. Department of Justice
Washington, D.C. 20537 <7>

Rev. LEAA 1975 ed. <8>
Revised by LEAA Data Archive and Research Support Center <9>
Center for Advanced Computation
University of Illinois
Urbana, IL 61801
(217) 333-3234

User's Guide Prepared by
LEAA Data Archive and Research Support Center <10>
(Under LEAA Grant 77-SS-99-6003)
December 1978 <11>

LEAA User's Guide 4th ed. <12>

Bibliographic Citation

Many MRDF which have value for secondary analysis are eventually turned over to data archives, but before this happens such data files are first cited in the published literature. Certain data producers rely primarily on the practice of scholars citing their data in the various research journals as the means of publicizing the existence of such MRDF. Consequently, the initial access route for many MRDF is via the cited reference in the research literature. The bibliographic elements outlined above for the title page are also used to create a citation or end-of-work reference. For example:

1. Author's full name
2. Title of data file: subtitle (if appropriate) [material designator][1]
3. Statements of responsibility (if appropriate)
4. City, State (abbreviated) of the data producer
5. Name of production organization [producer]
6. Date of production
7. City, State (abbreviated) of data distributor (if appropriate)
8. Name of distribution organization [distributor]
9. Notes (optional)

Item numbers refer to the bracketed numbers in the following example:

> <1> U. S. Law Enforcement Assistance Administration. <2> Juvenile Detention and Correctional Facility Census of 1971 [machine-readable data file]. <3> Conducted by the U. S. Bureau of the Census for the National Criminal Justice

---

1. The material designator is used to denote the generic form or type of material being referenced and to distinguish one type of medium from another. It is always enclosed in brackets and follows immediately after the title. Brackets are used to enclose the material designator, producer, and distributor statements.

Information and Statistics Service, LEAA, U. S. Department of Justice. LEAA rev. 1975 ed. Revised by LEAA Data Archive and Research Support Center, University of Illinois at Urbana. <4> Washington, D.C.: <5> U. S. Bureau of the Census [producer], <6> 1971. <7> Urbana, Ill: <8> LEAA Data Archive and Research Support Center [distributor].

Abstract

A bibliographic citation may also be used as a heading for the abstract. An abstract is an abbreviated and informative representation of the file being described. It is not intended to give information on a question by question level, but rather a summary of the major subject content. Its purpose is to tell the reader whether the file might be of interest and what is involved in obtaining it or securing more information.

Components of an abstract should include the following:

1. Unique identification numbers either for the abstract or study, if appropriate

2. Type of file (e.g., text, numerical, graphic, program source, etc.)

3. Bibliographic citation

4. Methodology:

   4.1 Source(s) of information

   4.2 Chronological coverage

   4.3 Universe description of target population

   4.4 Type of sample

   4.5 Instrumentation characteristics (e.g., telephone interview, mail questionnaire, etc.)

   4.6 Dates of data collection

5. Summary of the major subject content

   5.1 Purpose or scope of study

   5.2 Special characteristics of the study

   5.3 Subject matter

5.4 Number of variables, observations, and records.

6. Geographic coverage

7. Descriptors which express an idea of concept or phenomenon not covered in the body of the abstract (use terms that summarize the underlying conceptual framework of the study).

8. Technical notes

8.1 File structure (rectangular, hierarchical, etc.)

8.2 File size

8.3 Special formats (SPSS, SAS, etc.)

8.4 Computer or software dependence

9. Terms of availability

9.1 Condition of data (e.g., statements that edit checks have been made, etc.)

9.2 Restrictions on access, if any

9.3 Contact person or organization (full address and telephone number)

10. Cited references to any written or published reports which were based on these data and which might provide additional information for the potential user.

Example 2 (p. 17) shows an abstract.

EXAMPLE 2: ABSTRACT

Unique identification number(s): Accession number OP-003-004-USA-1957.

Citation: American Family Growth, 1957-1967. [Machine-readable data file]. Principal investigators, Charles F. Westoff et. al. DPLS ed. Edition prepared by Mary Ann Hanson, under the direction of Larry Bumpass, Center for Demography and Ecology, University of Wisconsin-Madison. Madison, Wis.: University of Wisconsin Center for Demography and Ecology [producer], 1978. Madison, Wis.: University of Wisconsin Data and Program Library Service [distributor].

Methodology. The target population was urban, native-born white couples with two children, couples whose marriages so far had been uncomplicated by death, divorce, separation, or extensive pregnancy wastage, with the second birth to have occurred during September 1956 for every couple. A probability sample, stratified by metropolitan area, was drawn from 7 SMSAs with population over 2 million (exclusive of Boston). Couples were interviewed three times in February-March 1957, 1960, and between 1963 and 1967 to determine eligibility and to complete questionnaires. Data checks and full-scale processing were run on the public use version. The final sample size is 1,165 couples; 814 couples completed all three interviews.

Summary of contents. American Family Growth, 1957-1967 is a longitudinal study which examines the fertility history of American couples in metropolitan America and the motivational connections between the environment and fertility decisions and behavior. Phase I looks at the social and psychological factors thought to relate to differences in fertility. Phase II focuses on why some couples stopped at two children while others had a third or fourth child during the first and second phase. Phase III examines how well attitudes and events of the early marriage determined the record of the later years of childbearing. The data file contains over 1000 variables.

Geographic coverage. United States SMSAs (New York, Indianapolis, Chicago, Los Angeles, Milwaukee, Cleveland, Minneapolis)

Descriptors. Fertility, family planning, family composition, socioeconomic status, work satisfaction, contraceptive practices, religiosity.

Technical notes. Rectangular file with 1,165 observations

Terms of availability. Data checks and full scale processing have been performed on the public use file. There

are no restrictions on access to the public use file. Copies of the data and documentation can be obtained by writing to the Data and Program Library Service, 4452 Social Science Building, University of Wisconsin-Madison, Madison, Wisconsin 53706 USA; telephone number: (608) 262-7962.

Cited references. Principal monographs include Family Growth in Metropolitan America by Charles F. Westoff, Robert G. Potter, Jr., Philip C. Sagi, and Elliot G. Mishler (Princeton, NJ: Princeton University Press, 1961); The Third Child: A Study in the Prediction of Fertility by Charles F. Westoff, Robert G. Potter, Jr., and Philip C. Sagi (Princeton, NJ: Princeton University Press, 1963); and The Later Years of Childbearing by Larry Bumpass and Charles F. Westoff (Princeton, NJ: Princeton University Press, 1970).

Example 2: Abstract

Pagination And Tables Of Contents

A User's Guide should use standard book pagination, with preliminaries in a lower case roman page sequence, and the body of the text and appendices in a single arabic page sequence. Two separate tables of contents are included. A table of contents for the entire document gives page numbers for preliminary and appendix sections, but has only a single entry giving the first page of the dictionary listing. A second table of contents contains an entry for each variable, together with headings for sections of variables.

Since a single edition of documentation is not likely to require extensive updating, there is little advantage in having separate pagination sequences for each appendix. Any extensive revision of the archival data set should result in the production of a new edition. Minor errors can be corrected by the publication of a replacement page. If a revision requires replacing a single page with more than one page, a decimal page increment such as 1.1, 1.2, can be adopted. However, such incremental paging should be avoided if possible.

Examples 3 and 4 (pp. 20 and 21) show tables of contents for the user's guide as a whole and for the data dictionary listing.

EXAMPLE 3: USER'S GUIDE TABLE OF CONTENTS

Abstract i
Data Dictionary Table of Contents v
Introduction 1
  Publications 1
  Scope of Data Collected 2
  Census Coverage 2
  Period Covered by the Census 3
  Data Collection Methods 3
  Archiving History 4
  Archive File Description 4
  Description of the Dictionary 5
Data Dictionary 7
Appendix 1: Definition of Terms 61
Appendix 2: Errors and Problems in the Data 64
Appendix 3: Counties by Name 67
Appendix 4: Questionnaire 73
Appendix 5: OSIRIS III Type 4 Dictionary Description 79
Appendix 6: Data Request Form 81

Example 3: User's Guide table of contents

EXAMPLE 4: DATA DICTIONARY TABLE OF CONTENTS

Identification Variables

Sequence Number 7
State Number 7
Level of Government 9
County Number 10
City Number 10
Check Digit 10

Agency Information

Type of Criminal Justice Agency 10
Type of Juvenile Facility 10
Agency ID Number 11
Sex of Population Served 11
Number of Employees 11
Resident Population as of June 30, 1971 11

Type of Facility or Program

Primary Function of Facility 12
Level of Government Administering Facility 13
Sex of Residents Held by Facility 13

Movement Data Reporting Period

Reporting Period Starting Date-Month 13
Reporting Period Starting Date-Day 13
Reporting Period Starting Date-Year 13
Reporting Period Ending Date-Month 14
Reporting Period Ending Date-Day 14
Reporting Period Ending Date-Year 14

Population Movement into Long-term Facilities

First Court Commitments-Male 14
First Court Commitments-Female 14
Recommitments by Court for New Offense-Male 15
Recommitments by Court for New Offense-Female 15
Returned for Violation of Aftercare/Parole-Male 15
Returned for Violation of Aftercare/Parole-Female 15
Returned from Aftercare/Parole for Other Reasons-Male ..15

Example 4: Data Dictionary Table of Contents

CHAPTER 3:
HISTORY OF THE ORIGINATING PROJECT

This section of the user's guide gives a short history of the data collection effort, including the rationale for the project, the reason for collecting these particular data, and, if appropriate, the statutory or executive authority for doing so. It is recommended that this section describe the evolution of the data from the original design of the study to the point where the data are converted to machine-readable format. The evolution of the data in machine-readable form is described in the section on file processing. The history section also includes the following items:

Instrumentation

The section on instrumentation describes the process by which information about a physical phenomenon was translated to numeric or coded form. The process of translation discussed in this section need not produce machine-readable information, but there should be a one-to-one correspondence between the information produced by the instrumentation and the data in machine-readable form. The history section should end at the point when data have been converted to machine-readable form. By convention, the section on file processing begins at the point immediately following the data's conversion to machine-readable form. The discussion here will be general, since it is intended to cover instrumentation in its general sense.

The instrumentation section must answer six questions:

1. What type of phenomena were measured? The answer to this question may seem obvious, but the user's guide should state explicitly that the data represent verbal responses, barometric pressures, geographic coordinates, etc.

2. Where, over what area, when, and over what period were the phenomena measured? An account of instrumentation includes discussions of populations, sampling, and estimation procedures, when appropriate.

3. What type of measurement or collection instrument was used? Where hardware instruments were used, give manufacturers, types, and model numbers. Give names, editions, and form numbers or versions of published or standard paper instruments and questionnaires.

4. How did the instrument operate? It is seldom necessary to describe the fundamentals of the instrument's operation. However, options and procedures specific to the study should be described. Such options include threshold and sensitivity settings for hardware, and probing instructions for interviewers.

5. How was the quality of the data monitored at the time they were collected?

6. What was done to correct data found to be in error?

These questions are not an outline for the the instrumentation section. Answers to some of them may take only a single phrase, while answers to others may require several labeled sections of text. In some cases, it may be useful to provide both a chronological account of the instrumentation and an account organized according to the questions above. For example, the quality of data is often monitored at the time of collection and again during the analysis of outliers. However, the analysis of outliers may take place several stages further along in the study.

Bibliography

The project history gives a bibliography for the project and ordering information for other documents relevant to the study or to the data file. A voluminous bibliography might better be included as an appendix to the user's guide.

An example of a project history (Example 5) follows on pages 24-26.

EXAMPLE 5: PROJECT HISTORY

For many years an annual survey of public facilities for adjudicated juveniles was conducted by the U. S. Department of Health, Education and Welfare (HEW) and published under the title, Statistics on Public Institutions for Delinquent Children (SPIDC).

The Juvenile Detention and Correctional Facility Census of 1971 represents the first census of such facilities operated by state and local governments in the juvenile justice system. (A second census, also held by the LEAA Data Archive, collected data for 1972 and 1973.)

The coverage of the present census has been broadened to include those public facilities which serve children awaiting court action as well as those already adjudicated. As a result, detention centers and shelters were included in the enumerated facilities, whereas previously only correctional facilities and diagnostic or reception centers had been surveyed. The addition of shelters and detention centers to the census completes coverage of publicly administered residential institutions in the juvenile justice system. The census was designed by LEAA and HEW, while the data collection effort was carried out by the Bureau of the Census.

Census Coverage

The census included juvenile detention and correctional facilities that were operated by state or local governments at the time the survey was conducted (October, 1971); these facilities had been in operation at least a month prior to June 30, 1971, and had a resident population of at least 50 percent juveniles. All facilities that held youthful offenders in addition to juveniles were also included. Juvenile detention centers that were part of adult jails were not included because their staff and budget figures could not be reported separately. An individual facility, such as a camp or annex, which was administratively dependent upon a parent institution, was counted as a separate facility if it was located in a separate geographic area. Nonresidential facilities, privately operated facilities, facilities exclusively for drug abusers or for dependent and neglected children, foster homes, and federal correctional facilities were not counted.

The 1971 census included 722 public juvenile detention and correctional facilities. They were classified as: (1) detention centers, (2) shelters, (3) reception or diagnostic centers, (4) training schools, (5) ranches, forestry camps, and farms, or (6) halfway houses and group homes.

Classification of the facilities was based on responses to the questionnaire, which asked the respondent to mark the type of facility most applicable according to the definitions provided (See Appendix 1).

Multi-functional facilities, such as training schools with reception centers or detention facilities with long-term treatment programs, were classified according to the function having the largest capacity or resident population.

Facilities administered by more than one level of government were classified according to the level of government providing the largest funding.

Period Covered By The Census

The census sought to cover the period July 1, 1970, through June 30, 1971. Institutional population data were collected for September 30, 1970; December 31, 1970; March 31, 1971; and June 30, 1971. Average daily population was computed from the populations on those four dates. Numbers of staff members were reported as of June 30, 1971.

Data regarding admissions and discharges as well as institutional costs were reported for varying reference periods. The majority reported for the period July 1, 1970, through June 30, 1971, as requested.

Data Collection Methods

In the summer of 1971, a mailing list of juvenile detention and correctional facilities was prepared using as a basic source, the National Criminal Justice Directory, compiled in 1970 by the Bureau of the Census for the Law Enforcement Assistance Administration. This directory list was updated from a number of other sources, including the mailing list maintained by HEW for the SPIDC; the 1970 Directory of Correctional Institutions and Agencies published by the American Correctional Association; the 1969 Master Facility Inventory maintained by the Bureau of the Census for the National Center for Health Statistics; the 1970 or the 1971 State Comprehensive Law Enforcement Plans for each State; the 1968 Directory of Juvenile Detention Centers published by The National Council on Crime and Delinquency; and the 1970 International Halfway House Association Directory. The updated list was then subdivided by state and sent to the juvenile correctional authorities in the respective states for review. The resulting list included 833 facilities, 111 of which were eliminated in the course of the census because they did not meet one or more of the coverage criteria.

Example 5: Project History

The census was conducted by mail, with an initial mailout in October 1971. Questionnaires were mailed to central agencies where this procedure had been used in the HEW study the previous year. Three hundred forty-seven questionnaires were mailed to 42 central reporters (34 State agencies and 8 local agencies). The remaining 486 questionnaires were mailed directly to facilities.

Facilities which failed to respond to the initial mailout were sent second and third mail requests and then telegrams if necessary. The response rate achieved was 100 percent for most data items. Telephone follow-up was used extensively to clarify inadequate and inconsistent survey returns.

Publications

Information about census coverage, census period, data collection, and the definitions of terms in the appendix were taken from the publication titled Children in Custody, prepared by the Statistics Division, National Criminal Justice Information and Statistics Service. This report is for sale by the Superintendent of Documents, U. S. Government Printing Office, Washington D. C., 20402, for $3.00 domestic postpaid. The stock number is 5-45-413-577.

In this publication, six facilities are misclassified; in addition some detention status data were incorrectly reported to the data collection agency. The data file has since been corrected for these errors. See Appendix 2 for details.

The reader is urged to read Children in Custody and this User's Guide carefully in order to understand fully the scope and limitations of the data. Anyone wishing to analyze the 1971, 1972, or 1973 data is referred to Appendix 2 of this User's Guide and to Juvenile Detention and Correctional Facility Census of 1972-73, also available from the LEAA Data Archive.

CHAPTER 4: FILE PROCESSING SUMMARY

The processing history section tells how the originating project or archiving facility edited the data into an archival file. This section describes general editing strategies rather than the corrections made to particular data items. Examples of editing strategies are eliminating unused high-order columns in data fields and changing missing data codes from alphabetic to numeric values. (Corrections of individual items should be listed in an appendix.)

Archive File Description

The file processing summary should give a detailed description of the machine-readable dictionary and data files, including the type of data management system on which they were prepared and the organization, number of records, number of variables, and record lengths of each file. In addition, this section should give the number of reels of tape required to distribute the file in various formats and recording densities.

Confidentiality Procedures

The file processing summary should contain an account of the assurances made to respondents concerning confidentiality and of any applicable statutes, regulations, or institutional requirements relating to the confidentiality of the data. The account should include references to use, maintenance, and/or disclosure of data, as appropriate.

The summary should also indicate whether the data are individually identifiable and, if not, whether and where keys to subject identifiers are maintained. Where data have been encrypted or otherwise modified in order to prevent identification of subjects, the file summary should indicate the nature of such procedures in order to insure that accurate conclusions are reached in the analysis of the modified data.

The file documentation should also contain a list of all variables that were deleted. If random values were added to variables, then the documentation should disclose the distribution functions and statistical parameters of the random values.

The section on confidentiality should have an explicit statement of the intent of the confidentiality procedures. Improvements in statistical techniques are making possible the reconstruction of individual information which was aggregated in order to protect respondents' privacy. The only reasonable way

of maintaining confidentiality is to restrict queries, rather than data. As long as data are released to a user who may make unrestricted queries, the data supplier cannot be sure that the data will not be disaggregated to a level below the supplier's intention. An explicit statement of the intention of the confidentiality-preserving procedures should inhibit the publication of results which violate assurances of confidentiality.

An example of a file processing summary (Example 6) follows on pages 29-30.

EXAMPLE 6: FILE PROCESSING SUMMARY

Archiving History

The LEAA Data Archive has done some editing, consistency checking, and correction of the data. Uncorrected errors, inconsistencies, and other known problems in the data are documented in Appendix 2.

The LEAA Data Archive ran frequency distributions for categorical variables and for minima and maxima for continuous variables in order to determine undefined or unreasonable values. Two-way cross-tabulations were run on the Governments Division of the Census Bureau identification variables and identification variables from the questionnaire where the variables seemed to represent the same information. The variables concerned covered the level of government administering the facility, type of facility, and sex of residents held. Other consistency checking concerned the existence of parole or aftercare programs. All known errors have been corrected.

Archive File Description

The archive file consists of three physical files stored on magnetic tape: a machine-readable User's Guide (this document) which documents the data for a human user, a machine-readable OSIRIS III Type 4 dictionary file which describes the data in a format readable by statistical programs, and a statistical data file. All information in the three files is in character format.

In the data file, the unit of observation is a juvenile facility. There are 722 units. Each unit is identified by variables, including state, county, and city numbers, level of government responsible for facility, type of agency, and an agency number. These variables were assigned by the Governments Division of the Bureau of the Census.

The records in the data file are sorted on: state number, level of government administering the facility, county number, city number, type of facility, and agency identifier (Reference Numbers 2-5, 8, and 9). A unique sequence number was assigned to each case in the sorted file.

All data in the data file are stored as numeric characters and are in whole numbers. The record for each facility is 544 characters in length and contains 245 variables.

The data and dictionary files were prepared using the OSIRIS III data management and analysis programs developed at the Center for Political Studies at the University of Michigan. The text in the User's Guide was formatted and printed using FORMAT, an implementation of FMT - A DOCUMENTATION PROGRAM, written by Bill Webb at the University of British Columbia.

In addition to the OSIRIS III system, for which the dictionary and data files were designed, they can be used directly with SPSS (Statistical Package for the Social Sciences), SAS (Statistical Analysis System), and MIDAS (Michigan Interactive Data Analysis System). The data can also be read into many other statistical packages. The necessary information about the format of the data can be found in the dictionary or in the dictionary listing section of this document.

Example 6: File Processing Summary

CHAPTER 5:
DATA DICTIONARY DESCRIPTION AND LISTING

The term "data dictionary listing" is used here to indicate the section of the user's guide giving the characteristics of each variable.

A data dictionary entry is a block of information pertaining to a single variable. The example on pages 34-35 gives the full syntax of the dictionary entry for a rectangular data file. However, there are some stylistic considerations which govern the construction of dictionary listing entries.

Dictionary Information

The dictionary information section of the entry contains all information from the file's machine-readable dictionary, or the information necessary to construct such a dictionary if none is provided. A dictionary information section is shown as items 1-8 in the example of a dictionary listing description. In general, dictionary information should be repeated in full for each variable, even if much of the information is constant over much or all of the file. The dictionary information section in the example included here contains all such information on each variable in the file.

Wording of questions. A codebook entry or series of entries of original variables should be as close as possible in wording to the original data collection instrument. However, in some cases clarity and brevity require that the original text be modified. In many cases, questionnaire items can be quoted in a user's guide exactly as they appeared on the original instrument. Questions originally in the form, "Do you own (or do) any of the following?" should be decomposed into a series of questions of the form "Do you own (do) ...?" This decomposition allows subsets of the documentation to be produced without rendering some items incomprehensible.

Where the original question text is quite long, it may be appropriate to place the bulk of the text in a paragraph at the head of a group of codebook entries. The paragraph can be included in any subset of the original documentation which includes variables to which the paragraph applies. In such cases, the question text for each variable need consist only of text differentiating that particular question.

Variable names or numbers. The variable name (or number) is not only an identifier but an address. This dictionary item is subject to the most stringent restrictions on its syntax. The dictionary information section must be explicit as to how

the variable is addressed. Some statistical systems, e.g., OSIRIS, allow only numeric variable addresses; others, e.g., SPSS, allow the variable address to be composed from a larger set of characters, including numeric and alphabetic.

Reference numbers. It is good policy to identify variables with a reference number as well as a variable name or number. The reference number is designed to maintain continuity in a variable's identity if its name or number is changed. Variable names may change when a file is read into a system with naming conventions different from the system on which the file was created. Variable numbers may be changed when files are subsetted or merged with other files. For this reason, all dictionary listing text refers to variables by a permanent reference number rather than by their current variable number. (While it would be possible to update all variable references in the user's guide dynamically, such a procedure would make it much more difficult to compare user's guides over different generations of an archival data set.)

Variable labels. The variable label (item 2 in Example 7) is the text other than the variable name or number used to identify the variable on printed output. Variable labels are subject to considerably more restrictions on length and character set than is question text. Thus, while it is important that text never be cryptic, it is often impossible to make the meaning of a variable label intuitively clear to the untutored reader. However, since the variable label is more a mnemonic than an explanation for the variable, cryptic variable labels can be tolerated. Extreme terseness can be tolerated in labels for a series of related variables, since the group of variables will provide a context for deciphering each member variable label. The main requisite for the label of a variable in a series is that the label identify the variable as being part of the series, and that it distinguish uniquely between it and all other variables in the series.

Extremely condensed variable labels may be composed by making liberal use of the arithmetic and logical symbols available in the ASCII and EBCDIC character sets.

Where a variable label cannot be sufficiently abbreviated by the use of symbols alone, it may be necessary to generate ad hoc abbreviations for words. Words at the right hand side of the variable label should be abbreviated first, with elimination of suffixes and vowels rather than of consonants. In some cases, the entire right-hand portion of a word may be left off without markedly decreasing the comprehensibility of the variable name. This strategy maximizes the amount of contextual information available to the reader scanning the variable label.

Explanatory text. Explanatory text is used to describe either a group of variables or a group of code values. An example of explanatory text for a group of variables would be a

paragraph containing the full text of a multiple response question which had been decomposed into a series of yes or no questions. Explanatory text, like all other references to variables, should refer to the catalog entries it describes with a permanent name.

Explanatory text can also be used to condense repetitive text of responses to questions, or as a subheading to classify a subset of responses. For instance, the values of a variable STATE might be grouped into categories such as "New England," "Midwest," etc.

Response text. A response section consists of a numeric code value, a short response category label, an optional extended response text item, and an optional frequency count. The abbreviated response text is used by several statistical systems to label printed output, and thus is subject to length restrictions. In the documentation system used in this manual, category labels are allowed 20 characters, but many statistical systems (e.g., SPSS) will print only the first 16 characters of the category label as table column labels. Where the original full text of the response must be shortened to fit the category label, the full response should be included as additional text following the category label.

Universe definition. Where a variable is not applicable to all observations, it is helpful for it to carry a logical expression defining the observations for which the variable is applicable. Using the example on pages 32-35 (Example 8), a variable which applied only to shelters and reception centers in California would have the statement,

UNIVERSE: V2=05 & (V3=2|3)

included after the question text.

Description Of Data Dictionary

The data dictionary description section is a set of annotated entries with detailed notes showing the function of each element. It is very important to insure that the data dictionary description include examples of all syntactic elements appearing in the user's guide. Since the production of an annotated data dictionary segment is quite tedious and prone to error, it may be wise to construct a standard example showing all possible options.

Examples 7 and 8 (p. 34 and pp. 36-39) show the user's guide's description of a hierarchical data dictionary and an example of a fragment of such a dictionary.

EXAMPLE 7: DESCRIPTION OF DATA DICTIONARY LISTING

This section contains an explanation of typical entries in the data dictionary listing section which follows. The numbers in brackets do not appear in the actual text, but are references to the descriptions which follow the examples.

.....

[1] V238
[2] FAMILY/JUVEN COUNSELNG?
[3] Location: 537 Width: 1
[4] Missing data: GE 3
[5] Implicit Decimal Places: 0
[6] Reference: 0238
[7] File I.D.: JD1
[8] Numeric character
[9] Record Type: 3

[10] VIII.D.1.(3) Does this facility routinely provide counseling involving the juvenile and his family?

[11]	[12]	[13]
413	1	Yes
309	2	No

[14] Ref. No. 239-245. These variables apply to training schools, forestry camps, and similar facilities, only. (REF 0213 = 4 OR 5)

.....

1. The variable number. The variable number is used as a variable name when the file is processed by systems (e.g., SPSS) which use alphabetic names.

2. The abbreviated variable label used by statistical systems to identify the variable on program output.

3. The starting location and width of this variable in the data as stored on a magnetic tape. This is the format information needed to read the data into other systems.

4. The designation of missing data. In the example, code values for FAMILY/JUVEN COUNSELNG? greater than or equal to 3 (GE 3) are missing data. Other notation, EQ meaning equal to or LE meaning less than or equal to, may also be used. Many data management and analysis packages require that certain types of data which are usually excluded from analysis be designated as "missing data"; e.g., inappropriate, unascertained, unascertainable or ambiguous data categories. If need be, this designation can be changed by the user and the values used in a substantive manner.

5. The number of decimal places which are implicit in the variable. A decimal point actually written into the data item will override any implicit decimal place entry in the dictionary. (This item need not be included when there are no implicit decimal places. It has been inserted here for illustrative purposes.)

6. A reference number. Both a variable number and a reference number are assigned to each data item when a file is created. In the archive file, they are the same. Should the data then be subsetted or rearranged, the variable numbers may change to reflect the order of the new data file, while the reference numbers remain to reflect the variable numbers in the dictionary listing describing the archived data file.

7. The three-column identification number unique to this data file. The file identification and the reference number together should uniquely identify the variable in any context.

8. The storage mode of the variable. All data in this file are in numeric character form.

9. The type identifier of the record in which the variable occurs. This information need not appear for rectangular files, which have only one type of record.

10. The unabbreviated variable label or the text of the question from the survey form. The letters and numbers at the beginning of the text are the question numbers from the survey form. Although in many instances the actual question text has been paraphrased, an effort was made to use the same vocabulary and to retain the meaning of the original phrasing. Dependency and contingency of a variable is indicated here when appropriate.

11. The frequency of occurrence of each code value.

12. The code values occurring in the data for a variable.

13. The textual definitions of the codes. The first 20 characters form a short label which some systems use to document the output of analysis programs. A longer description follows the short label when necessary.

14. Explanatory text referring to a group of variables which follows. These entries are used whenever a lengthy description applies to more than one or two variables or when some comment about a group of variables is needed.

Example 7: Description of Data Dictionary Listing

EXAMPLE 8: DATA DICTIONARY LISTING

Record Identification Variables

V0 Reference: 0000
RECORD TYPE IDENTIFIER File I.D.: JD1
Location: 1 Width: 1 Numeric character
<No Missing Data Defined>

51 1 STATE

722 2 FACILITY

Record Type 1, State

V101 Reference: 0101
SEQUENCE NUMBER File I.D.: JD1
Location: 2 Width: 3 Numeric character
<No Missing Data Defined> Record Type: 1

Sequence numbers were assigned to all records in the file after facilities records were merged into state records. Sequence numbers range from 1 to 773, for 51 state records and 722 facility records.

V102 Reference: 0102
STATE NUMBER File I.D.: JD1
Location: 5 Width: 2 Numeric character
<No Missing Data Defined> Record Type: 1

This variable keys Record Type 1, State, to Record Type 2, Facility.

01 Alabama

(V102, STATE NUMBER, Continued)

02 Alaska

03 Arizona

04 Arkansas

05 California

V103 Reference: 0103
TL STATE FACILITIES File I.D.: JD1
Location: 7 Width: 3 Numeric character
Missing Data: EQ 999 OR GE 998 Record Type: 1

Total number of juvenile detention and correction facilities in the state.

Record Type 2, Facility

V201 Reference: 0201
SEQUENCE NUMBER File I.D.: JD1
Location: 2 Width: 3 Numeric character
<No Missing Data Defined> Record Type: 2

The values range from 1 to 773 and are in ascending order. Sequence numbers were assigned to the file sorted on Ref. Nos. 202, 203, 204, 205, 208, and 209.

Ref. Nos. 202-212. These variables are the identification variables assigned by the Governments Division of the Bureau of the Census.

V202 Reference: 0202
STATE NUMBER File I.D.: JD1
Location: 5 Width: 2 Numeric character
Missing Data: EQ 99 Record Type: 2

This variable keys Record Type 1, State, to Record Type 2, Facility.

9 01 Alabama

Example 8: Data Dictionary Listing

(V202, STATE NUMBER, Continued)

3 02 Alaska

9 03 Arizona

7 04 Arkansas

105 05 California

V213
PRIMARY FUNCT OF FACILTY
Location: 21 Width: 1
Missing Data: EQ 9

Reference: 0213
File I.D.: JD1
Numeric character
Record Type: 2

II.A. What is the primary function of this facility?

305 1 Detention center
Provides temporary care in a physically restricting facility for juveniles in custody pending court disposition, and often for juveniles who are adjudicated delinquent or are awaiting return to another jurisdiction.

17 2 Shelter
Provides temporary care, similar to that of a detention center, in a physically unrestricting facility.

16 3 Reception/diagn cntr
Reception or diagnostic center. A facility that screens juvenile court commitments and assigns them to appropriate treatment facilities.

191 4 Training school
A specialized institution serving delinquent juveniles committed directly to it by juvenile courts or placed in it by an agency having such authority.

115 5 Ranch/fors camp/farm
Ranch, forestry camp, farm. A residential treatment facility for juveniles, whose behavior does not necessitate the strict confinement of a training school, often allowing them greater contact with the community.

78 6 Halfway house
Halfway house, group home. A facility where children live in the facility but are permitted extensive contact with the community such as for jobs and schools.

Example 8: Data Dictionary Listing

V290	Reference: 0290
AGE OF OLDEST MALE	File I.D.: JD1
Location: 233 Width: 2	Numeric character
Missing Data: EQ 99	Record Type: 2

IV.A.2. Age of the oldest male currently under supervision. Warning: See Appendix 2 for more information about this variable.

Example 8: Data Dictionary Listing

CHAPTER 6: APPENDICES

A user's guide need not have any appendices. Most often, appendices are used either for bulky information which would otherwise break up the continuity of the main section of the user's guide, or for information about the current state of the data file, which may change over the file's use.

Definitions

Where the subject matter of a file includes unusual terminology, or where terms have special definitions relevant to the data, the documentation should include an appendix of definitions. An example of a section of a definition appendix follows on page 41 (Example 9).

EXAMPLE 9: APPENDIX: DEFINITIONS OF TERMS

Following is a glossary of terms, concepts, and categories used in collecting the Juvenile Detention and Correctional Facility Census of 1971 . This glossary was, for the most part, taken from the above-cited publication, Children in Custody.

Classification Of Residents

In terms of a person's being charged with a criminal offense, a juvenile is one over whom the juvenile court has original jurisdiction in cases of delinquency. The juvenile court's jurisdiction is determined by the age of the client who must, in most states, be under 18 years of age. In this census, the actual definition of a juvenile or child was left to each jurisdiction, since no universal definition seemed applicable to all phases of the individual's contact with the juvenile justice system.

A juvenile who has been adjudicated delinquent is one who, through formal judicial proceedings, has been adjudged guilty of a criminal offense or has been declared in need of supervision by the court. Purely for statistical purposes, voluntary admissions to juvenile facilities were also tallied as adjudicated delinquents. Voluntary admissions include juveniles who committed themselves or who were referred to the facility for treatment by parents, court, school or social agency without being adjudged delinquent or declared in need of supervision by a court.

Juveniles awaiting transfer to another jurisdiction are juveniles who have allegedly committed a crime in or have run away from another jurisdictional area, including runaways from correctional facilities. Juveniles adjudicated delinquent and awaiting placement in a correctional facility are not included here but in the "juveniles adjudicated delinquent" category.

Juveniles held pending disposition by a court are juveniles held for delinquency who have not had any hearing or who have had only a preliminary hearing or screening, and who are awaiting further court action.

Errors And Problems

When it is not possible to correct all known errors in a data file, known errors should be listed in an appendix to the user's guide. Where appropriate, errors are listed for individual cases. An example of an appendix listing known errors and problems is shown on pages 43-44 (Example 10).

EXAMPLE 10: APPENDIX: ERRORS AND PROBLEMS IN THE DATA

This appendix contains a discussion of known errors, inconsistencies, and other limitations in the 1971 juvenile facility census data.

In the first section, errors are listed in order by variable, the affected cases are identified, and the status of corrections noted. Also, the dictionary listing section contains a warning and a reference to this appendix in the entry for each affected variable.

The second section contains a discussion of limitations on the comparison of the 1971, 1972, and 1973 data.

Some of this information was generated by the LEAA Data Archive during the process of archiving the data file. Other information was obtained from Ms. Jenny Eldreth, National Criminal Justice Information and Statistics Service, LEAA (202) 376-2622.

Errors By Variable And Case

The case identification variables for this file are:

Ref. No. 2: STATE NUMBER
Ref. No. 3: LEVEL OF GOVERNMENT
Ref. No. 4: COUNTY NUMBER
Ref. No. 5: CITY NUMBER
Ref. No. 6: CHECK DIGIT
Ref. No. 7: AGENCY TYPE
Ref. No. 8: TYPE OF FACILITY
Ref. No. 9: AGENCY ID NUMBER

Ref. No. 90: AGE OF OLDEST MALE

Ref. No.	2	3	4	5	6	7	8	9
Case 1	39	0	021	000	0	5	4	19

Problem: Value of 0052 (age 52 years) looks too high.

Status: This answer has been verified on the original questionnaire form. The institution held adults as well as juveniles.

Use And Comparison Of 1971, 1972, And 1973 Data

Before comparing any of the three years for which data are available, the user should be aware of several general restrictions.

The 1972 data must be used with caution because they were collected in late 1973 as part of the combined 1972-73 census. This arrangement meant that facilities in existence in 1972 but not in late 1973 could not be represented in the data. In addition, some respondents could not provide actual data or reasonable estimates. LEAA considers the 1972 data to be minimum counts and chose not to compare them with data for any other years.

The categories of movement and population data for 1971 and 1973 are not exactly the same, so that some judgment is necessary in comparing these years. For the final report of the 1972-73 census, an estimating procedure was used to inflate the 1973 movement data and make them comparable with the 1971 data.

The 1971 detention status category "adjudicated delinquent" includes persons in need of supervision and voluntary commitments. These categories are reported separately in the 1973 data and must be summed before comparing with the 1971 data.

The same kind of problem occurs in the staffing data. In 1973, payroll staff was reported separately from non-payroll staff and both terms were defined on the questionnaire form. In 1971, the questionnaire requested counts of "employees" but in fact the data includes both "employees" on the payroll and other staff not on the payroll. In 1973, non-payroll staff included community volunteers who were not counted in 1971.

In 1971, definitions of full-time and part-time were stated in the questionnaire; they were not stated in the 1973 questionnaire. Also, there is no certainty that the data were reported as requested in 1971. Because of these reporting discrepancies, LEAA has chosen to compare only full-time staff.

To compare 1971 full-time staff with the 1973 data, the 1973 payroll and non-payroll staff must be summed. Also, the 1971 data classified staff into specific job titles. The 1973 data include more general categories. For these data, the 1971 data must be summed in order to permit comparisons with the 1973.

Example 10: Appendix: Errors and Problems in the Data

Extended Response Categories

Where a variable has a very large number of response categories, it may be helpful to put the full listing of response categories in an appendix so that the user need not constantly flip through page after page of text in order to get through the dictionary listing. In addition, the use of appendices for bulky ancillary material makes it possible to split the user's guide into two volumes, one of which contains the heavily used dictionary listing.

Examples of such codes include the complete list of more than 3,000 counties in the United States, or the Census Bureau's coded lists of occupations and industries. Since all the entries in this appendix are in the data dictionary, no action is needed to insure that county labels are printed in the course of analyses. The user may, of course, simply insert the appendix into the dictionary listing in the proper place if desired.

An example of part of an extended response category listing follows on page 46 (Example 11).

EXAMPLE 11: APPENDIX: EXTENDED CODE CATEGORIES

The list which follows contains the name of each county for which there is a record in the data file. The list includes and is ordered by the state number and the county number, Ref. Nos. 2 and 4 respectively.

01 - Alabama

037 - Jefferson
045 - Madison
049 - Mobile
051 - Montgomery
052 - Morgan

02 - Alaska

026 - Ketchikan, Gateway
027 - Greater Anchorage
032 - Matanuska, Susitna

03 - Arizona

001 - Apache
002 - Cochise
003 - Coconino
004 - Gila
007 - Maricopa
008 - Mohave
010 - Pima

04 - Arkansas

004 - Benton
010 - Clark
026 - Garland
035 - Jefferson
060 - Pulaski

05 - California

001 - Alameda
003 - Amador
004 - Butte
007 - Contra Costa
008 - Del Norte
010 - Fresno
012 - Humboldt
013 - Imperial
015 - Kern
019 - Los Angeles
020 - Madera
021 - Marin
022 - Mariposa
023 - Mendocino
024 - Merced
027 - Monterey
028 - Napa
029 - Nevada
030 - Orange
031 - Placer
033 - Riverside
034 - Sacramento
035 - San Benito
036 - San Bernardino
037 - San Diego
038 - San Francisco
039 - San Joaquin
040 - San Luis Obispo
041 - San Mateo
042 - Santa Barbara
043 - Santa Clara
044 - Santa Cruz
045 - Shasta
047 - Siskiyou
048 - Solano
049 - Sonoma
050 - Stanislaus
052 - Tehama
054 - Tulare
056 - Ventura
057 - Yolo
058 - Yuba

06 - Colorado

001 - Adams
003 - Arapahoe
016 - Denver
021 - El Paso
028 - Huerfano

Cross References To Earlier Editions

Where an archival data file has gone through several editions, it may be appropriate to include a cross reference showing the different numbers assigned to a variable in different editions of the data file, or in a longitudinal data file. The dictionary listing example included here (Example 7) contains no such cross reference.

Original Data Collection Instrument

The user's guide should include a copy of the original data collection instrument as an appendix. If possible, one of the original questionnaires should be bound into or included with the user's guide. If no original is available, then a photographically reproduced document should be included. Only as a last resort should a totally reprinted copy of the instrument be produced for inclusion in the user's guide.

Field Or Laboratory Procedures

Include either a complete copy or (where voluminous) a sample of the field or laboratory procedures documents sufficient to allow an understanding of what transpired when the data were collected.

Dictionary Format

Just as the dictionary listing gives a complete format for the data file, the documentation should include a complete format description of the machine-readable dictionary. Even though the processing history section of the introduction gives the name and version of the data management or statistical system which generated the data file, it cannot be assumed that data users will have access to the original system, to its successors, or to their documentation. An example of a dictionary format description follows on page 48-49 (Example 12).

EXAMPLE 12: APPENDIX: OSIRIS III TYPE 4 DICTIONARY

Dictionary-descriptor Record[2]

Position	Content
1-3	Blanks
4	4: Type 4 dictionary
5-8	First variable number
9-12	Last variable number
13-16	1: Number of logical records per case
20	1: Format of variable location specification is starting location and field width
21-80	Blanks

Variable-descriptor Record

Position	Content
1	T: Variable descriptor record
2-5	Variable number
6	Blank
7-30	Variable name
31	Blank
32-35	Starting location of the variable within each data record
36-39	Field width of response 1-9: numeric variables 1-255: alphabetic variables
40	Number of decimal places 0-9: numeric variables 0: alphabetic variables
41	Character type and storage mode Blank: numeric, character mode 1: alphabetic, character mode
42	Variable type

---

2. This description of an OSIRIS III Type 4 dictionary was taken from OSIRIS III, Volume 1: System and Program Description, Release 2 Edition, Appendix D, published by Institute for Social Research, University of Michigan.

	Blank: single response 1: multiple response
43-44	Number of responses If the variable has more than one response, columns 32-39 describe the location of the first response and the remaining responses are assumed to be in adjacent fields of the same description.
45-51	First missing data code or blanks
52-58	Second missing data code or blanks
59-72	Blanks
73-75	File ID
76-80	Sequence number or blanks

Example 12: Appendix: OSIRIS III Type 4 Dictionary

Description Of Physical Shipment

This appendix contains a physical and logical description of the file and accompanying documentation as shipped by the data archive. An example of a shipment description is found on page 51 (Example 13).

EXAMPLE 13: APPENDIX: DATA SHIPMENT DESCRIPTION

The 1971 juvenile facility data, second edition, are distributed on magnetic tape in three files: the machine-readable user's guide, the dictionary, and the data. Tape-recording specifications and a partial listing of the tape in internal format are also provided.

For additional copies of this document, the user's guide file can be listed on a line printer. Each record is 133 characters long and contains an ASA carriage control character in column one. The text is in upper and lower case, and, although it was intended for an IBM TN print train, none of the special TN train characters were used. There are approximately 3,400 records in the file, which is named JD71.XAJ.

The dictionary, an OSIRIS III Type 4 dictionary file, contains 246 records of length 80. The complete format is described in Appendix 5. All information is in character format. The file name is JD71.3AJ.

The data file contains 722 records of 544 characters each. All data are in numeric character format, and the file name is JD71.AAJ.

To order a copy of the data please complete the form on the following page and mail it to:

Ms. Barbara Noble
LEAA Data Archive
Center for Advanced Computation
University of Illinois
Urbana, IL 61801
(217) 333-7164

Order Form

The user's guide should contain an order form specifying the ways in which data may be supplied. The form should allow the client to choose alternatives, rather than requiring the writing in of technical information such as recording densities and blocking factors.

An example of an order form appears on page 53 (Example 14).

EXAMPLE 14: DATA ORDER FORM

Name: _____

Address: _____

_____

_____

Tape Recording Specifications

Seven-track Tape

Density (BPI)	200	556	800
Parity	Even	Odd	
Record blocking	Blocked	Unblocked	
Maximum block size	_____		
IBM standard labels	Labeled	Not labeled	
1-6 character label	_____		
Character code	BCD		

Nine-track Tape

Density (BPI)	800	1600	6250
Parity	Odd		
Record blocking	Blocked	Unblocked	
Maximum block size	_____		
IBM standard labels	Labeled	Not labeled	
1-6 character label	_____		
Character code	EBCDIC		

(Over)

Example 14: Data Order Form

CHAPTER 7:
DOCUMENTING MACHINE-READABLE DATA: A CHECKLIST

This checklist summarizes the file documentation standards of this manual. It is designed to be used in evaluating the formatting and documentation of machine-readable data files.

1. Is there a title page which includes the following elements?

   1.1 Title (Subtitle, if appropriate) which describes

      1.1.1 File contents (subject matter)

      1.1.2 Geographic focus

      1.1.3 Chronological year(s) of data target or data collection

      1.1.4 Study or series identification (when applicable)

   1.2 Authorship

   1.3 Special sponsorship or funding source

   1.4 Producer's name and address

   1.5 Date of production or collection

   1.6 Edition statement (if appropriate) for data

   1.7 Distributor's name, address, and telephone number

   1.8 Sponsorship statement for documentation (if different from data)

   1.9 Date of documentation's publication

   1.10 Edition statement (if appropriate) for documentation

   1.11 Library classification code(s)

   1.12 Catalog facsimile

1.13 Copyright (if relevant)

1.14 If there is no title page, can one be constructed from other information in in the user's guide?

2. Is there a data acknowledgements statement?

3. Is there a study description in which the following elements appear?

3.1 Bibliographic information

3.1.1 Author

3.1.2 Title

3.1.3 Medium designator

3.1.4 Edition

3.1.5 Imprint:
Date of production
Name of producer
Place of production
Date of distribution
Name of distributor
Place of distribution

3.1.6 Collation

3.1.7 Notes (optional)

3.2 Descriptors or key words

3.3 Subject matter (logical contents)

3.4 Data collection history

3.5 Instrumentation

3.6 File processing history

3.7 Physical organization (structure)

3.8 Conditions of availability

3.9 References (key publications)

3.10 If there is no study description, can one be constructed from information in the user's guide?

4. Is there a Table of Contents to the user's guide?

5. Is there a Table of Contents to the "dictionary" or codebook?

6. Is there a history of the collection project in which the following elements appear?

   6.1 Organization and objectives (as appropriate)

      6.1.1 Legislative or executive mandate

      6.1.2 Grant or contract number

      6.1.3 GSA and OMB reporting codes

   6.2 Overview of data collection methods and analysis

   6.3 Description of sampling design/selection

      6.3.1 Universe

      6.3.2 Sample design

      6.3.3 Units of analysis

      6.3.4 Weighting

      6.3.5 Response rates/replacements

      6.3.6 Panels/replications

      6.3.7 Sampling reliability (or statement of unreliability)

   6.4 Instrumentation

      6.4.1 Type of collection instrument(s)

      6.4.2 Time, place, and duration of data collection

      6.4.3 Operation of collection instrumentation

      6.4.4 Audit and quality control of collection

      6.4.5 Correction and update procedures

   6.5 Coding procedures/coding error rates

6.5.1 Manual edit procedures

6.5.2 Coding errors

6.6 If the user's guide contains no project history, did it cite reports, articles, or other publications giving such information?

6.7 Document availability:

6.7.1 Included in bibliography

6.7.2 Accompanied user's guide

6.7.3 Available in a library, special information center, or through a publishing house

6.7.4 Available only from the project staff

7. Is there a file processing history in which the following elements appear?

7.1 Description of conversion to machine-readable form

7.1.1 Description of types of records in file

7.1.2 Logical relations between record types

7.2 Global consistency checks

7.3 Recodes and creation of new variables

7.4 Missing data conventions

7.5 Procedures for protecting confidentiality

7.6 If the documentation includes no file processing history, did the user's guide indicate information available in reports, articles, or publications?

7.7 What documentation of the data is available?

7.7.1 Accompanied user's guide

7.7.2 Available in a library, special information center, or through a publishing house

7.7.3 Available only from the project staff

7.8 Is there a machine-readable data dictionary?

7.9 What was the data management system used to prepare the data described?

7.10 Is there a description of the physical structure (number of records, variables, record lengths of each file, format and recording density)?

8. Is there a data dictionary in which the following elements appear?

8.1 Variable number

8.2 Abbreviated variable name

8.3 Variable name

8.4 Reference number(s)

8.5 Explanatory notes or comments (other attributes of variables, such as units or dimensions, validity, etc.)

8.6 Abbreviated value labels

8.7 Complete text of explanations for each value of a discrete variable

8.8 Valid range

8.9 Frequencies or percentages of values (for discrete variables), summary statistics (for continuous variables)

8.10 Missing data representation scheme

8.11 Physical location in file

8.12 Width

8.13 Implied or coded number of decimal places

8.14 Storage mode (E.g., character, binary, packed decimal, etc.)

8.15 Number of responses, if a multiple response variable

8.16 Dependency or contingency

8.17 Record type in which variable is located

9. Are their appendices? If so, which elements appear?

9.1 Definitions (glossary of terms, concepts and categories)

9.2 Descriptions of unavailable data

9.3 Descriptions of errors and problems identified in processing and analysis

9.4 Coding schemes (Standard, well-known, authoritative?)

9.5 Cross references to earlier editions or files in a series (if appropriate)

9.6 Indices and measurement tools

9.7 Copy (facsimile) of original data collection instruments

9.8 Copies of laboratory or field procedures manuals and documents

9.9 Copy of coding instructions

9.10 Description of special methodological or design problems

9.11 Complete format description of machine-readable dictionary (if appropriate)

9.12 Data and documentation shipment description

9.13 Order form

9.14 Document availability:

9.14.1 Included in bibliography

9.14.2 Available in a library, special information center, or through a publishing house

9.14.3 Available only from the project staff

10. Organization of Documentation

10.1 Does user's guide describe process by which data produced in a logical, orderly sequence?

10.2 Is the text well written?

10.3 Are descriptions brief?

10.4 Is the text legible?

10.5 Does the text format ease retrieval of information (e.g., important points are underscored, new sections are color coded)?

10.6 Would a user need help in organizing the documentation?

10.7 Could a user begin immediate review of the documentation?

10.8 Can the documentation be used without professional assistance?

10.9 Is the documentation well organized?

10.10 If not, are there inhibiting factors?

10.10.1 Materials need organization

10.10.2 Size is too large

10.10.3 Materials too complex (require expertise in organization of information)

10.11 Not in hard copy medium (microfiche, magnetic tape, other medium?)

10.12 Materials incomplete?

11. Physical Characteristics

11.1 Is there a statement of tape characteristics?

11.2 Is the tape density and recording format compatible with available equipment?

11.3 Is the character set or storage mode of the tape compatible with that of available equipment?

11.4 Is the physical record length of the data within the reading capacity of available equipment?

11.5 Is there a listing of the tape's table of contents and a dump of selected files and records?

12. Logical Characteristics

12.1 Is the record identified?

12.1.1 Unique number

12.1.2 Record type

12.2 Do the data correspond to the documentation?

12.2.1 Size of record

12.2.2 Number of records

12.2.3 Number of record types

12.2.4 Number of records of each type

12.2.5 Number of files

12.3 Is identification complete?

12.3.1 All variables identified

12.3.2 All categorical values of variables identified

12.4 Are field widths sufficient?

12.5 Are computational and noncomputational variables identified?

12.5.1 Noncomputational variables apply to names and labels (never used in an arithmetic or numerical operation)

12.6 Explicit missing data value (values)

12.7 Standardized codes

12.7.1 Geographic

12.7.2 Occupation and job history (includes Dictionary of Occupational Titles, occupation and industry classification schemes)

12.7.3 Socioeconomic status

12.7.4 Political parties

12.7.5 Educational institutions

12.7.6 Social-psychological measures

12.7.7 Occupational measures

12.7.8 Political measures

12.7.9 Other

12.8 Are data at lowest level of aggregation:

12.9 All variables?

12.10 Some variables?

12.11 No variables?

12.12 Are all data from the collection included in the data file?

12.13 If all data are not included, can they be obtained?

12.13.1 Confidentiality rules prohibit access

12.13.2 Established by principal investigator/project staff

12.13.3 Established by sponsor

12.13.4 No reason

13. Bibliographic Practices

13.1 Were good bibliographic practices followed?

13.2 Unique identification of the product

13.3 Bibliographic Reference

14. Information Dissemination Channels

14.1 Were information channels used to disseminate the data? If so, which channels?

14.2 Government channels

14.2.1 Government publications

14.2.2 Government Printing office

14.2.3 National Technical Information Service

14.2.4 National Archives and Records Service

14.3 Scholarly journals and newsletters

14.4 Centers for dissemination of data

14.5 Multiple channels

14.6 Do announcements/publicity describe the data file? If so, which elements appear?

14.6.1 Full bibliographic references

14.6.2 Contents of the file

14.6.3 Relevant publications

14.6.4 Release conditions

14.6.5 Retention status (if appropriate)

14.6.6 Mode of access

14.6.7 Condition of the data

14.6.8 File exchange information

14.6.9 Contact person and telephone number

14.7 Do publications based on the data carry a notice identifying the data file?

15. Archival File Preparation

15.1 Were a public use data file and documentation prepared? If so, is the public use version the complete micro-level data (original) file?

15.2 If not, are reasons cited for their unavailability?

15.2.1 Confidentiality

15.2.2 Large size of file

15.2.3 Distribution constraints

15.2.4 Lack of resources

15.2.5 Nonstandard file format

15.2.6 Lacks adequate documentation

15.3 Does the distributor inform the user community of updates, changes and modifications in the data and documentation? If so, what channels are used?

15.3.1 Special mailings

15.3.2 Professional journals and publications

15.3.3 User requests

15.4 Does the distributor offer special services related to the data file? If so, what services are offered?

15.4.1 Subsets/extracts to specification

15.4.2 Performs analysis

15.4.3 Consults on analysis

15.4.4 Supplies related reports on analysis of the data (maintains a library)

16. Archival Arrangements

16.1 Were original (micro-level)data archived? If so, by whom?

16.1.1 Principal investigator(s)

16.1.2 Funding agency/sponsor

16.1.3 Producer

16.1.4 Distributor (not an archive)

16.1.5 National Technical Information Service

16.1.6 Library

16.1.7 Other

16.2 Data archive

16.2.1 National Archives and Records Service

16.2.2 Other federal agency designated as archive for data

16.2.3 Academic data archive

16.2.4 Other

16.3 Are data available from more than one source? If so, which sources?

16.3.1 Government and academic

16.3.2 Government and private

CHAPTER 8:
TECHNICAL STANDARDS FOR MACHINE-READABLE DATA FILES

This chapter sets forth some suggested standards of good practice for the formatting of machine-readable data files. It is assumed throughout this chapter that the physical medium of the file is magnetic tape. However, most of the material in the chapter is appropriate to machine-readable data files on any medium.

Tape Recording Standards

Data should usually be shipped on seven or nine track magnetic tape written at a density of 800, 1600, or 6250 BPI. The tape should carry IBM or ANSI volume and file labels. The best format for keeping archive tapes is IBM format 6250 BPI Group Encoded tapes. Not only does this density allow for more data per tape, but group encoding is more reliable than is the phase encoding technique used at 1600 BPI, which in turn is more reliable than the "non return to zero" (NRZI) technique used at 800 BPI and below. These reliability considerations make it advisable to ship at the highest density for which both shipper and receiver have equipment, even when the data fit easily on the tape.

Whenever possible, the tape volume and its files should be labeled with IBM or ANSI labels. While some installations may have difficulty in reading the labels, the number of computers which are unable to handle labeled tapes is declining rapidly. Without volume and file labels, there is always a possibility of a tape's identity's being lost when it becomes separated from its documentation. IBM or ANSI user labels can also be used to provide documentation for a machine-readable data file. Whenever possible, the user should be given the choice of IBM, ANSI, or no labels on a transmission tape. However, it is good professional practice for the data supplier to suggest that any label is better than no label. [3]

---

3. User header and trailer labels are 80-character records of text which are written onto the tape after the file label records. User labels are ignored by the operating system, but can be printed with a utility program.

Where files consist of more than one record type and are of variable length, it is good policy to allow users the option of a file of fixed length records formed by padding all records to the length of the longest record in the file. Files should be blocked to a length which will provide for efficient handling of the data. Physical record lengths of from 80 to 2000 characters are usually acceptable. If possible, users should be allowed to state the largest physical record size which they can conveniently accept. [4]

A set of volume and file naming conventions should be worked out early in the life of an archive or data producing project. While it is impractical to devise a set of file-naming conventions which are universally applicable, it will often be in the interest of data users and suppliers to establish naming conventions for sets of data files.

Data Types

Archived machine-readable data records should consist entirely of alphanumeric, EBCDIC or ASCII characters. (Binary length fields for variable length records are not included in this restriction.)

For the purpose of this chapter, variables will be considered either computational or noncomputational. Noncomputational items contain information such as names and labels, which are never used in any arithmetic or numerical operation.

Computational items contain numeric information which is designed to be used in computations. Noncomputational variables may contain any EBCDIC or ASCII printing characters. Noncomputational items should be left justified and padded to the right with alphabetic blanks. Computational variables may contain only the characters 0-9, ".", "+", and "-". The only exception to this is where D- or E-format floating-point data

---

4. Data on magnetic tape are divided into physical and logical records. A logical record contains data from a single "unit" (e.g., defendant, gas bill, library catalog card, etc.). Data processing is often more efficient if several logical records are written end-to-end onto the tape to form a single "physical" record. This process is called "blocking", and a physical record is sometimes called a "block".

are represented. [5] It is acceptable to represent floating point data in integer format, with implicit decimal places noted in the documentation.

Computational variables should be right justified and padded to the left with zeros or blanks. If a computational field is signed, the sign should immediately precede the left-most numeric character. Computational fields should contain at least one numeric character. In particular, computational fields consisting only of a sign or of blanks should not be permitted, even for systems which read blanks as zero. Fields containing only a signed zero are not acceptable, since some computers cannot represent a signed zero.

Missing data. All variables for which there may be missing data should have an explicit missing data value or values. In particular, it should never be assumed that a value of zero or a field consisting entirely of blanks indicates missing data. [6] Missing data values may be indicated in the documentation as a list or a range of values, and may include zero, but should not include values of -0, a blank field, or a field consisting only of a sign.

Missing data values should occur in the same field as the variable to which they refer. If an alternate value is to be used in place of a missing data value, the base item should carry an appropriate missing data code, while the alternate value is shown in a separate variable which has been declared for that purpose. In no case should the alternate value be carried in the base variable, with an explanation code in another variable. The rationale for this standard is that the meaning of a variable should be determinable without reference to a second variable. If the true value of a variable has been suppressed or modified, then the value of the variable should indicate such suppression or modification. If an alternate value is to be offered in such cases, the appropriate variable in the record can then be read if the analyst so wishes.

5. Floating-point data are used to represent very large or very small numbers. The number, 2,625,000,000 is written in floating-point notation as "2.625 E+09". The "E+09" is called the exponent, and indicates that the decimal point is to be shifted nine places to the right. D-format floating-point notation is similar, but uses a "D" rather than an "E" to mark the exponent. The number "0.00000000000465" is written in floating-point notation as "4.65 D-12". The "D" indicates that the number was computed with double precision arithmetic, which allows more significant digits to be represented.

6. It is acceptable for the documentation or data dictionary accompanying the file to say that missing data are represented by blanks. It is unacceptable to require the user to make this assumption in the absence of file documentation.

File Organization

Data items. Archived machine-readable data files should contain no undocumented or irrelevant fields. The width of a variable should be sufficient to accommodate the entire range of variation which may be expected of the item, but should not be excessive. For instance, a file containing data collected between 1960 and 1965 need allow only one character to indicate the year of collection. In general, the width of a variable should allow for variation in the high-order digit, rather than requiring imputation of the high-order digit. Thus a file containing data collected between 1958 and 1963 should have a two-character "year of collection" item, rather than requiring the analyst to imply the change of decade.

Conversely, fields need be no wider than is required to accommodate the maximum expected value of a variable. Thus, the variable "number of offenders" for a multiple victimization incident probably need not be larger than two characters, and certainly no larger than three. In any case, the maximum allowable field width for an integer variable should be nine or fewer characters. Larger numbers should be represented by supplying a scaling factor in the documentation, or by the use of floating point format. [7]

Record type identification. Each record in a file containing more than one record type, i.e., a hierarchical file, should carry an explicit type identifier. Even where each record in a file has a different length, an explicit type identifier should be included. A different record type identifier should be used whenever there is a significant change in any of the procedures used to generate that record. Such identifier changes may also be necessary for record layout, instrument design, data collection, and coding. The need for a new record type is obvious when coding instructions or code values are changed. However, even when such changes consist only of the addition of new coding categories to existing variables, a new record type should be produced. Otherwise, analysts may not properly interpret the presence or absence of variables which do not occur in all types of records.

For example, several minor changes were made in the coding categories and format of the incident record of the National

---

7. A scaling factor allows fewer characters to be used in representing a variable. For instance, $50,000,000 can be represented as "50" if the variable is defined as "Dollars in millions."

Crime Survey Victimization file. In order to interpret a "type 1" record properly, it is necessary to know what year the data in the record represent. A better method would be to include the record type and collection period when generating a set of identifiers for incident records, being sure to identify the particular format used.

Record identification items. Each record in a file should carry an identification number which is unique to the data file. If no existing variable will suffice as a unique identifier, then a sequence number should be assigned by the data supplier. Where a file contains more than one type of record, each record should carry an explicit type identifier, as well as a unique sequence number. Where the records in a file represent a hierarchy or tree, a record should carry identification sufficient to uniquely identify it and its position in the hierarchy. In particular, it should not be necessary to infer the location of a record in a hierarchy solely from its position in the file.

For example, consider a file consisting of household, person, and incident records. Each record in the file should begin with the same four data items: A record type indicator, a household identifier, a person identifier, and an incident identifier. An incident record will carry a type and an incident identifier, as well as the identifiers of the person and household to which it belongs. A person record will carry a type identifier, the identifier of the household to which it belongs, its own identifier, and a dummy incident identifier. Record identifiers should be positive integers, while dummy identifiers should be fields of zeros. Negative numbers and blanks should be avoided as dummy identifiers.

Most structured files represent a tree, a hierarchical structure with a single base element. Some files, however, may consist of records arranged as a lattice, a hierarchical structure in which an element may be connected to one or more of of several base nodes. (An example would be a file of records on individual children, each of which is related to several classroom records, and to a family record.) Records of files representing lattices should carry sets of identifiers sufficient to identify uniquely their position in the lattice.

In general, identifiers of lower level records need be unique only within level, since concatenating identification

variables generates a unique identifier. [8] In some files, there will be more than one type of record at the same level of the hierarchy. Questions as to whether sequence numbers should be unique within record type or within levels should be resolved before the file is generated.

The rationale underlying the assignment of a unique identifier to each record is that users of a data file should be able to perform arbitrary sorts and to subset the file without requiring the use of any facilities other than a sort program and a file-copying utility.

Dates. One problem confronting the analyst is ascertaining the period represented by a data file when it is part of an ongoing series of data collections. This is one of the problems of the Bureau of the Census' Current Population Surveys, where the studies are carried out on an almost monthly basis and questions are repeated yearly (e.g., the Annual Demographic File). Where a file is one of a time series of identically structured files, the date should be treated as one of the record identifications. In such cases, it is helpful if the date is represented as YYMMDD, where YY is the year, MM the month, and DD the day of the month. This order facilitates sorting by date.

Standardization Of Data Codes

Where possible, standard data codes should be used. While there are no universal standards, it is often possible to find a widely used set of codes which is appropriate. Geographic coding can be done either with FIPS (Federal Information Processing Standards) codes or with Census Bureau codes. The choice of which "standard" coding scheme is used is not so important as that ad hoc coding schemes are used as little as possible. Data producers who have created what they believe to be generally useful coding schemes should so indicate in the documentation accompanying a data file. Data archives should investigate claims of generally useful coding schemes and should disseminate such schemes to their suppliers and users.

---

8. To concatenate identifiers is to string them into a single identifier. Consider a file of households with people in them. Suppose that each household has a three-digit identifier which is unique within the file, and each person has a two-digit identifier which is unique within the household. Then person number 4 in household number 207 is uniquely identified as person number 20704.

Documentation

Each tape should be accompanied by documentation giving the physical characteristics of the volume and files, as well as the logical composition of each type of record. Wherever possible, the documentation should be in machine-readable form and supplied both in hard copy and as a file on the tape. Machine-readable documentation is preferred because it prevents the separation of documentation from data and because the physical quality of the documentation will not be degraded by repeated copying.

Machine-readable documentation should be in printer image. If a document processing program has been used to format source text, it would be helpful if the source text (and the name of the processor) were included with the printer image documentation. [9]

Minimal documentation. The minimal documentation of a data file consists of a tape volume table of contents, a character format listing, and octal or hexadecimal (internal format) listing of a sample of records of each file, and a minimal codebook. The name, title, and affiliation of the principal investigator responsible for collection of the data submitted, the source of funding (including the grant number, if any) should be included in the documentation. Where appropriate, data files should be accompanied by copies of the original collection instruments, including survey questionnaires and interview schedules. Copies of editing and coding instructions used in creating the data file should also be included.

Tape table of contents. Each tape transmitted to a user should be accompanied by a tape table of contents generated from that particular volume. The tape volume table of contents listing should include all information from the volume label and from the file labels. Information should be in an easy-to-read form, rather than an internal format listing of the text of the labels.

If possible, the tape table of contents should be produced by a program which also verifies the readability of the tape. If no tape listing program is available, then the tape table of contents should be produced manually, using information from the job which produced the tape or an internal format listing of the

---

9. The current flux in word processing techniques makes it impractical to attempt to standardize source text or document processing languages at this time.

tape. [10] An example of a Tape Table of Contents is found on page 74 (Example 15).

10. Since tape listing programs are becoming increasingly available, it is suggested that data centers without such programs attempt to acquire them.

EXAMPLE 15: TAPE VOLUME TABLE OF CONTENTS

TAPE NAME = *RSW009* 24 MAY 1977 12:20:06

IBM-LABELED 6250-BPI 9TP VOLUME=RSW009 OWNER=CAC,U-ILL RACK#=C4524
LP=ON BLK=ON RING=OUT DTCHK=ON RETRY=10

FILE #	DATA SET NAME	BLOCK COUNT	RECORD COUNT	TAPELTH (FEET)	RECORD FORMAT	BLOCK LTH AV.	BLOCK LTH MAX.	CREATED DD MMM YY	EXPIRES DD MMM YY	USER I.D.	BATCH RECEIPT#
1	NCS73.NAT.CQ1	1233	125858	283.07	VB(15250,305)	15141	15250	23 MAR'77		SGDA	
2	NCS73.NAT.CQ2	1234	126442	283.17	VB(15250,305)	15135	15250	23 MAR'77		SGDA	
3	NCS73.NAT.CQ3	1187	123279	272.60	VB(15250,305)	15146	15250	23 MAR'77		SGDA	691127
4	NCS73.NAT.CQ4	1212	125436	278.21	VB(15250,305)	15139	15250	23 MAR'77		SGDA	691127
5	NCS74.NAT.CQ1	1086	112135	249.39	VB(15250,305)	15139	15250	23 MAR'77		SGDA	691127
6	NCS74.NAT.CQ2	1083	112086	248.84	VB(15250,305)	15148	15250	23 MAR'77		SGDA	691127
7	NCS74.NAT.CQ3	1072	111090	246.18	VB(15250,305)	15138	15250	23 MAR'77		SGDA	691127
8	NCS74.NAT.CQ4	1100	113455	252.62	VB(15250,305)	15141	15250	23 MAR'77		SGDA	691127

TOTAL TAPE LENGTH = 2114.07 FEET

<*><*><*> END OF TAPE <*><*><*>

81

Example 15: Tape volume Table of Conte

Minimal Codebook

The codebook included with the file should contain at least the following information for each variable:

1. A reference number.
2. An unambiguous name for the item.
3. A textual description of the item, or the text of the question, if from a questionnaire.
4. The starting location, width, location of implicit decimal point, or scale factor.
5. Missing data codes and their meanings.
6. The mode in which the variable is represented, i. e., numeric character, alphanumeric string, floating-point binary, etc.

The codebook should also contain a list of the valid values for categorical items, and valid ranges for continuous items. Missing data codes should be documented in the same fashion as other values, and not left implicit.

Frequency tables. A frequency distribution for each categorical variable should accompany each file. The mean, standard deviation, range, and number of cases of continuous variables should also accompany the file. Values which fall outside of those defined in the codebook should be annotated if they cannot be corrected.

*U S GOVERNMENT PRINTING OFFICE: 1980 311-378/1412